DOCUMENT RESUME

ED 079 346 TM 002 945

AUTHOR	Langmuir, Charles R.
TITLE	Cross-Validation.
INSTITUTION	Psychological Corp., New York, N.Y.
REPORT NO	Bull-47
PUB DATE	Sep 54
NOTE	4p.; Reprint from Test Service Bulletin
JOURNAL CIT	Test Service Bulletin; n47 p3-6 Sep 1954
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bulletins; *Item Analysis; *Statistical Analysis; *Test Selection; *Test Validity
IDENTIFIERS	*Cross Validation

ABSTRACT

Cross-validation in relation to choosing the best tests and selecting the best items in tests is discussed. Cross-validation demonstrated whether a decision derived from one set of data is truly effective when this decision is applied to another independent, but relevant, sample of people. Cross-validation is particularly important after statistical data have been used to choose the best tests to make up a battery for use with the next group of people. A cross-validation experiment on a new group will tell how good the choice of test really is. Item analysis is a means of improving tests. The data from the analysis are used to eliminate doubtful items and to determine the best scoring weights. By applying the revised test to a new independent group, the inventory is refined. When the test is, without change, administered to an entirely new and independent set of criterion groups, cross validation data are obtained. (For related document, see TM 002 947.) (DB)

Test Service Bulletin

No. 47 THE PSYCHOLOGICAL CORPORATION September, 1954

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to 304 East 45th Street, New York 17, N. Y.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

JEROME E. DOPPELT
Assistant Director

DOROTHY M. CLENDENEN
Assistant Director

ALEXANDER G. WESMAN
Associate Director of the Test Division

JAMES H. RICKS, JR.
Assistant Director

ESTHER R. HOLLIS
Advisory Service

CROSS-VALIDATION

PEOPLE keep asking us, "What's this talk about cross-validation?" Perhaps this is a good time to explain what we think the jargon is all about. In the simplest language we think cross-validation means taking another independent look, especially verifying a first choice or checking up on a hunch. The idea seems to us to be hoary with age. At least the notion of taking a second look was well established in horse-and-buggy days. The driver, you remember, was cautioned at every grade crossing to Stop, Look, and Listen. Fancy language attaches to such a primitive notion only because the complexities of choosing the best tests for some purpose and selecting the best items in test construction introduce special difficulties.

The problem of cross-validation is the problem of getting an independent verification, and the special difficulties we have in our work of personnel testing arise in our need to select the items that look best and the tests that look most useful from a large number of possibilities. We believe in the experimental approach. We like to try out tests and items and choose the "best" after seeing the results. Statistics – especially correlation coefficients and item analysis statistics – get into the play.

We really have two problems. The first is to find the right way to choose the "best" of a number of possibilities. The second is to find out how good our best choice actually is. Cross-validation is concerned principally with the second problem.

We often think in this connection of one of our bright-eyed friends, Angelo, our barber around the corner. He thinks he has learned quite a bit about statistics from us and as a matter of a fact he has. Greatly to his sorrow, he cross-validated a selection system by experimental verification, using horses as subjects. But that is the way things are in this life. Empirical studies and particularly cross-validation have an insidious way of destroying confidence in systems of evaluation and prediction. Our friend Angelo, however, is a persistent fellow. He has given up the ponies, but he is still looking for the practical gimmick. He is now working on a fascinating plan, guaranteed by logic, wholly objective and backed up, he says, by extensive cross-validation statistics.

He starts with a pool of one hundred thousand equivalent items, each one as comparable to the others as the pennies in the piggy bank. In fact the items *are* Lincoln pennies, and the price is about as economical as an item can be. Angelo has set up a simple scheme for administering and scoring these items. He flips a coin and scores the obverse side (heads, to you) Republican. Of course the reverse side is classified Democrat. (Angelo denies that the fact that Lincoln *was* a Republican introduces bias in the key. He says the key is arbitrary and he will change it if anyone insists.) His theory is that the pennies that can predict (really post-dict) the 1900 election will be "good" items. He has already tried out the 100,000 coins, and approximately 50,000 of them turned out to be scored

Republican. This was for the year that Roosevelt (Teddy) was elected, and these 50,000 coins clearly called the election.

The 50,000 discriminating pennies look like good ones for betting purposes. But Angelo's experience with horses has convinced him that cross-validation is in order. He plans to repeat the experiment for the election year 1901 to identify the coins that were good enough for the 1900 study but not good enough for 1904. This process is called skimming the cream of the item pool. Since there is no limit to the extent to which one can improve tests by careful screening of the items, Angelo's plan is to extend the process right down to the election of 1952. He expects to wind up with half a dozen strictly comparable forms of the very best test for the job of predicting elections, and he thinks they will be pretty valuable. Extended sequential experiments will have proved that these are the pennies that have successfully predicted fourteen consecutive elections. Angelo is sure that he will be able to retire in November 1956.

Chance is the gremlin. The larger the number of predictors — be they tests, items, or pennies — the more careful we must be to guard against being fooled by chance results that may "look" meaningful. Angelo should have realized that the 50,000 pennies he picked in his experimental try-out for 1900 had no special virtue except that of the true impartiality with which chance endows all honest, two-sided pennies. And in every successive election year when he "skimmed the cream of his item pool" or "purified" his scoring system he made the same fundamental error.

• • • • •

We don't know of any dictionary that defines cross-validation as psychologists use the term. Perhaps we can clarify the idea with examples.

Suppose we find a test that is reputed to be a good selector of salesmen. We need such a test in the worst way and hope this is it. But we know that a man's success at selling vacuum cleaners door-to-door may be no good for predicting how well he will do at selling a line of tools to hardware stores. So we try out the test in our client's line of selling, being careful to test an adequate number of applicants and to keep the test scores put away where they won't influence anyone until the performance records come in for checking. This is a *validation* experiment. An appropriate statistical analysis will show whether the test scores correlate with our criteria of success as well as we expect from the evidence which led us to try the test. This type of study might well be called *more validation* rather than *cross-validation.* If we apply the same test to several similar groups such as samples of salesmen in similar work and find that the several validity coefficients are about the same, we can have *more confidence* in using the test for selecting this general class of salesmen. Mosier, in the reference cited later, uses the expression "validity generalization" for this kind of result.

Studies with the *Differential Aptitude Tests* provide abundant examples. Page 42 of the manual lists 36 coefficients of correlation between Numerical Ability scores and subsequent course grades in mathematics in several schools and classes. The coefficients range from .27 to .65; the median r is .47. In a loose sense the data contribute to cross-validation, but it is clearer to think of such correlation studies as examples of validity generalization. Many such studies in a variety of situations are required before a generality of confidence in test validity can arise in the mind of the careful test user.* The notion of cross-validation ties in to the general validity question, but it is more specific to particular applications of tests.

After we have selected some tests for practical use we usually set some cutoff score for each test or some combination of scores on several tests which will maximally eliminate potential failures and maximally include potential successes. We try to decide on appropriate cutoff scores or weightings of scores by studying the data on a sample of candidates. When we apply these *decisions* to a new sample of similar candidates we are ready to cross-validate our findings. That is, we are ready to take a second look at the rules we decided on. If the cutoff or weighting system shows up well we have accomplished a favorable cross-validation and probably will adopt the system. However, the results may not be as good as we expected. In this case the cross-validation study is negative. The results warn us that more research is necessary. The essence of the idea is that cross-validation demonstrates whether a *decision* derived from one set of data is truly effective when *this decision* is applied to another independent, but relevant, sample of people.

Cross-validation is particularly important *after* we have used statistical data to choose the best tests to make up a battery for use with the next crop of applicants. A cross-validation experiment on a new group

*See *Test Service Bulletins* 37 and 38 for discussion of validation, especially on the need for many studies.

will tell us how good our choice really is. The purpose of cross-validation is to protect us from being fooled into putting confidence in a relationship which *happens* to hold true for the group we started with, but which will let us down in the long run. And we don't get this protection unless we make sure:

(a) that the scoring system and combination of tests picked on the first group is tried out unchanged on the second;

(b) that the second group is a relevant sample of different people.

• • • • •

Suppose that we have several hundred items sampling personality attributes, interests, or attitudes; and suppose further that we have paired criterion groups, say, successful and unsuccessful salesmen, or male and female sophomores, or bank presidents and clerks. Assume, too, that factors such as age and education have been controlled. Since we do not know what responses characterize the groups and especially do not know what responses identify the significant traits of the different groups, we administer the items as a test to the groups and perform a detailed item analysis. We are simply seeking empirically those items that are usefully discriminating. The analysis provides a method of identifying some of the characteristics that distinguish between the types of person. The data are used to revise the experimental inventory, to guide the elimination of poor items (or at least those that look poor) and perhaps to determine the best scoring weights (or at least to determine what look like the best scoring weights).

After we have identified them we could score all the selected items according to the empirically determined key and see how well the scores differentiate the original groups or correlate with the criterion. This work would doubtless encourage us to press forward to publication of our apparently important findings, but being careful workers we realize that cross-validation is desirable. We may also argue that the criterion groups are small and we are sure that the validity might be improved by further refinements. Let's try it again, we say. So let us suppose that at great cost and effort, we are successful in obtaining new independent criterion groups. We now apply the revised test to these groups.

It is obvious that every test can be improved, and in the process of analyzing the second sample we note that some of the items are apparently miskeyed, or the scoring weights seem wrong. We plan, therefore, in the statistical work to include an independent second item analysis. Surely two item analyses are better than one, and we have only to find the best method of evaluating the results of the double item analysis to achieve greater validity. If we have enough items we may eliminate all the doubtful ones, namely those in which the second item analysis fails to confirm fully the original findings. These items, we would say, are of doubtful validity, and if we have enough items, we prefer to discard them in favor of items that have been doubly proved or twice-validated. So we rescore all the tests on both samples for the reduced number of items in the second revision and with the refined key. We obtain scores and compute appropriate coefficients of correlation for the second independent sample and the original sample separately and for the two groups combined.

All this labor produces an impressive pile of data, and we may think we have cross-validated our test and our weighting system. *As a matter of fact, we have only refined the inventory. We have as yet no validation data for the revised instruments.*

We do not have cross-validation data until we administer the tests without change, without further revision or refinement, to an entirely new and independent set of criterion groups.

Now there is nothing wrong in trying out items on a number of samples. The more objective data we have the better should be our judgments about items to include. But when we have finally put the items together and developed a scoring system, we should undertake a new validation study completely independent of the samples used in the developmental phases of the work. A published "validity" coefficient based on the sample which contributed to the selection of the items and the making of the key (in the case of personality and interest inventories) is misleading. Coefficients so derived should be unambiguously described. They are *not* validity coefficients which tell the practical user what he may expect if he uses the test or inventory.

In this connection we recall an amusing experimental example. The experiment is simple and quite instructive concerning the way fortuitous accidents of sampling can affect the selection of items for an inventory. It happened that we had a conference of school people, ten high school principals and ten superintendents. To illustrate the point about cross-validation we offered to build right then and there what we call the Wardrobe Projective Inventory for Administrative Personnel. The test is designed to dis-

tinguish between principals and superintendents. The underlying psychological construct in our test development is obviously expressed in the well-established truth, "Clothes make the man."

This is the way to apply the theory: ask each member of the conference committee to answer about 50 or 60 simple check-list items describing his wardrobe of the day. For example, is his suit blue, grey or brown? Single breasted or double breasted? Shoes – black, brown, two-tone? Necktie – four-in-hand or bow, quiet or loud? Shirt – white, colored; plain or button-down collar; French or plain cuffs or half sleeves? And how about the socks? Wool, silk, cotton, orlon, dacron, nylon, plain or fancy? Get the facts by such a check list and then let the statistician analyze the data. He will find which items distinguish the principals from the superintendents. He counts the frequency with which each answer characterizes each group. If we but try enough items some of them will surely turn out to look significant in distinguishing between the men with different jobs.

It turned out in our data that the questions about suits, shoes and neckties taken as single items did not exhibit any useful significance. About the same number of superintendents and principals chose brown suits and shoes and bow ties and white shirts. But the combination two-toned shoes, bow tie and pastel shirt with button-down collar was significantly a superintendent's choice. Not a single principal chose the combination. By further careful analysis of the frequency counts for combinations of items we chose the wardrobe check-list items of greatest differential significance. We derived from the data sufficient experimental insight to construct a Superintendent-Principal key. The scores distinguished with 90% accuracy whether a man *in this group* was a superintendent or principal. The question is: Is this really a valid measuring device?

Though we have not tried to duplicate the experiment we are sure on logical grounds that we could find discriminating items in a new group of principals and superintendents. But the set of questions that would look "significant" on the second study would include different items.

Duplication of the experimental process of item selection on independent samples of people would not be cross-validation. True cross-validation would be a trial of the selected items on new groups. If we apply the principle of independence to the subsequent trials the cross-validation data will tell us how well the test really works. We surely expect that a trial of the Wardrobe test on independent groups of principals and superintendents would reveal truthfully the non-significance of the items of the check list. We will have caught up with the gremlins of chance.

We have emphasized the test maker's primary responsibility for cross-validating the selection and weighting of test items to produce good psychometric instruments. The same fundamental ideas also apply to the test consumer's responsibility for cross-validating his use of tests in specific practical applications. Whenever a variety of measuring devices is tried experimentally to guide the choice of the most valid battery some one of the tests will correlate best with the criterion. By appropriate statistical methods a combination of tests may be weighted to yield the greatest multiple correlation. It is certain that fortuitous accidents of sampling have influenced the choice of tests and the weighting of the scores. The chance effects may be very important – so important that repetition of the experiment might result in a different choice of tests or a very different regression equation. A true estimate of the value of a weighting system, regression equation, or a choice of cutoff scores on a single test should be derived by cross-validation in the test user's actual application.– *Charles R. Langmuir.*

NOTE: Some of the interesting ramifications of the problem are discussed in a symposium, "The Need and Means of Cross-Validation," by C. I. Mosier, E. E. Cureton, R. A. Katzell, and R. J. Wherry in *Educational and Psychological Measurement*, Vol. 11, No. 1, Spring 1951, and a classic experimental example is reported in "Validity, Reliability, and Baloney," by E. E. Cureton in the same journal, Vol. 10, No. 1, Spring 1950.